Natural Language Processing and automatic SNOMED-encoding of free text:
An analysis of free text data from a routine Electronic Patient Record application with a parsing tool using the German SNOMED II

Joerg H. Hohnloser •, Matthias Holzer *, Martin R.G. Fischer *, Joseph Ingenerf**, Alexandra Günther-Sutherland*

• Institut f. Med. Informationsverarbeitung, Biometrie and Epidemiologie
* Med. Klinik, Klinikum Innenstadt
Ludwig-Maximilians-Universität München, Germany
**GSF-Forschungszentrum für Umwelt und Gesundheit, Medis-Institut, Neuherberg, Germany

A significant proportion of data entered into medical computer systems ist still free text format. The flexibility of expression and ability to express uncertainty are offset by difficulties to automatically process this information. We have developed a Natural Language Processing tool including automatic SNOMED II [1] coding (German version, [2]). We did not modify or expand the SNOMED vocabulary through any thesaurus. The tool enabled us to parse free text sections of our Electronic Patient Record System PADS [3]. PADS has been up and running in an acute care environment since 1991. We specifically analyzed the section on "chief complaint" and "reason for admission". We present data on two different sets of questions answered by the NLP parser:

1.) Was the patient admitted because of symptoms suggestive of acute myocardial infarction ?

2.) Did the patient admitted have an operative Procedure ? What type of procedure was done ? Which organ system was the subject of the procedure ?

For each group of questions a random sample from a database of 7000 admissions was selected. In order to be eligible the free text sections needed to contain information on the relevant questions. The NLP results were then evaluated by an experienced acute care physician manually processing every single record thereby serving as the gold standard. Furthermore, for each question we evaluated 20 records as a control group which did not contain any information relevant to the target question areas. For both sets of questions 40 records plus 20 controls were processed, respectively. Our results show a specificity of 100 % and a mean sensitivity of 59 % (46.3 % for procedural terms, 70.4 % for diagnosis terms). If a procedural term was identified, the exact procedural technique and organ system involved were correctly coded in 88.3 and 71.4 %, respectively. 17 concepts for question 1 (0 for question set 2) were "recognized" which clearly were artefacts of the NLP-parser and had no corresponding free text information in the records. The reasons for thę artefacts were related to the NLP-parsers approach to dissect a medical term into syllables which occasionally contained fragments of other SNOMED terms. This was done to increase sensitivity to terms nested in other compound terms. The term fragments were then recognized by the parser which reproduced new (and inappropriate) SNOMED terms. In this version of SNOMED (v. II, German) modifier terms (such as status post ..., rule out ... etc.) and typical abbreviations for medical terms were not included to the degree necessary for reasonable parsing. Furthermore, one problem unsolved is the terminology overlap between abbreviations used to describe medical terms and the same abbreviations used for other terms such as chemical compounds (i.e. metals) or lab tests.

We conclude: Our NLP parser for the German SNOMED II version can be used to process large volumes of aggregate free text records for purposes of rough quantification and qualification of problems. Specificity and sensitivity can be adjusted but will affect each other unfavourably. The tool is useful to identify target records otherwise not accessible to traditional database query techniqes because of their free text nature and complexitiy of compound termsto generate a medical "meaning". This, however, requires specific questions to be asked in order to not include artefacts produced in the parsing process.

The tool should not be used on a case to case basis or in situations where high sensitivity and specificity are needed in combination. Our data suggest also, that the additon of a clinical thesaurus would definitely add value to such a system.

Key Words:
Natural language processing, free text, electronic patient record system PADS, German SNOMED II

1. Rothwell, D. J., and Hause, L. L. SNOMED and microcomputers in anatomic pathology. Med Inf (Lond) 1983; 8: 23-311

2. Wingert, F. SNOMED- Systematische Nomenklatur der Medizin. 1984 Berlin: Springer-Verlag

3. Hohnloser, J.H., Pürner, F. (1992) PADS (Patient Archiving and Documentation System): A computerized patient pecord with educational aspects. International Journal of Clinical Monitoring and Computing 9:71-84